Credit Risk Assessment: Estimating Probability of Default using Logistic Regression

Ritesh Parbhoo

2026-02-08

Email: riteshparbhoo@outlook.com

GitHub: https://github.com/ritesh-parbhoo/

Project Overview

Project Description

This project, implemented in R, involves developing a credit risk model to estimate the probability of default of borrowers using logistic regression (GLM). The model examines how key variables - employment status, bank balance and annual salary - predict the probability of a counterparty defaulting on their credit obligations. The model is trained on a training set and tested on a validation set, and various methods are used to evaluate the discriminatory power of the model.

Motivation

Credit risk - the risk that a counterparty will default on their financial obligations - is a principal concern of lenders and financial institutions. Key to measuring credit risk is estimating probability of default which will allow institutions to make data-informed lending decisions, manage risk exposure, and comply with regulatory requirements. By developing a logistic regression model using key predictors - employment status, bank balance and annual salary - this project aims identify high risk borrowers, thus improving credit allocation and reducing potential financial losses.

```
library(pROC)
library(PRROC)
library(ggplot2)
library(caret)
```

Data Overview

About the Dataset

The data.frame containing the entire dataset is labeled df. The response variable, labeled "default_status" is a binary factor taking on values 0 or 1. The predictors consist of another binary factor, "employment_status", and two numerical values, "bank_balance" and "annual_salary", which have both been standardised and stored in the data.frame as "st_bank_balance" and "st_annual_salary" respectively.

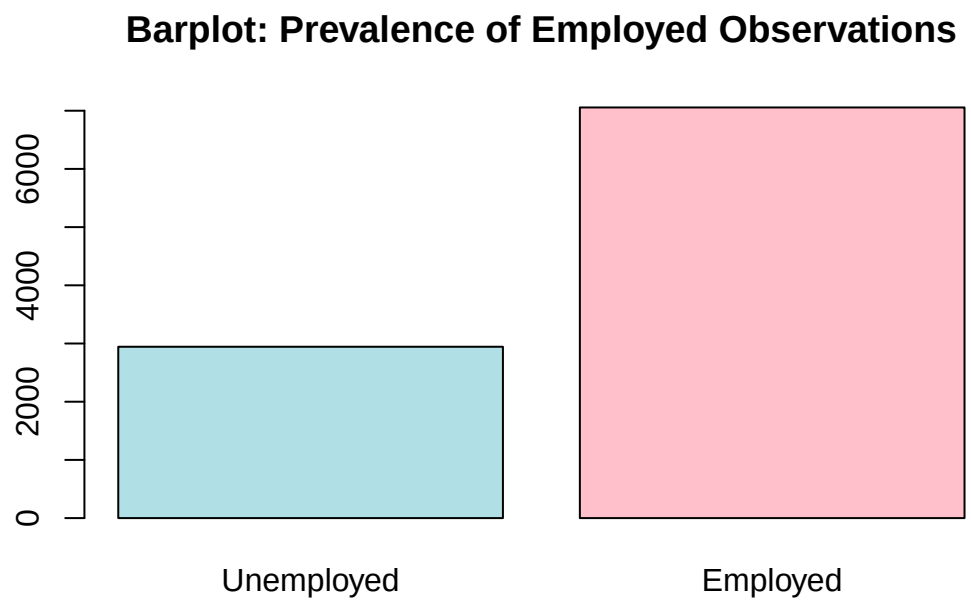
```r
print(head(df))
```

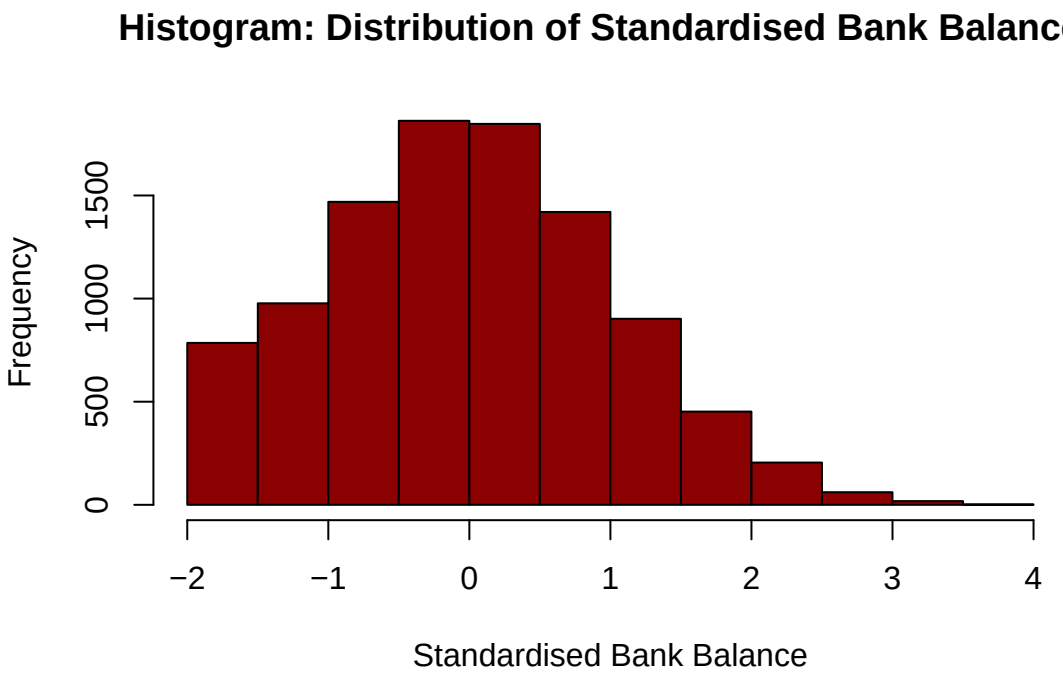
```
##   employment_status st_bank_balance st_annual_salary default_status
## 1                 1     -0.21881663        0.8131470              0
## 2                 0     -0.03761487       -1.6054158              0
## 3                 1      0.49238734       -0.1312056              0
## 4                 1     -0.63286213        0.1640224              0
## 5                 1     -0.10277722        0.3708969              0
## 6                 0      0.17410072       -1.9514227              0
```

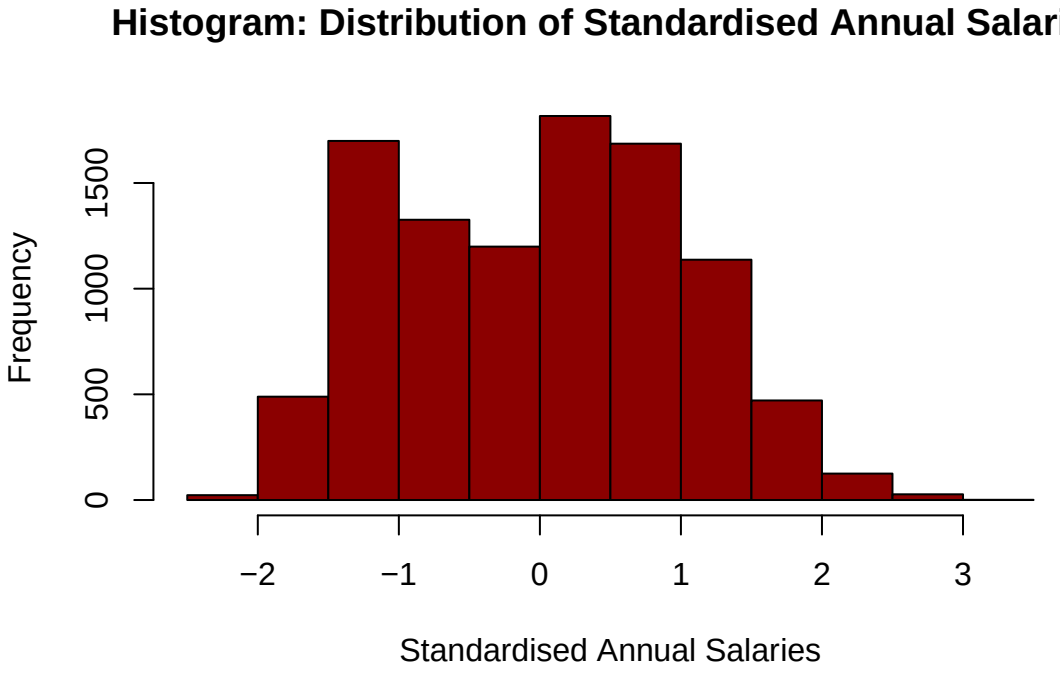
```r
print(summary(df))
```

```
##  employment_status st_bank_balance    st_annual_salary   default_status
##  0:2944            Min.   :-1.72700   Min.   :-2.45527   0:9667
##  1:7056            1st Qu.:-0.73110   1st Qu.:-0.91301   1: 333
##                    Median :-0.02427   Median : 0.07766
##                    Mean   : 0.00000   Mean   : 0.00000
##                    3rd Qu.: 0.68414   3rd Qu.: 0.77161
##                    Max.   : 3.76037   Max.   : 3.00205
```

Visualising the Data



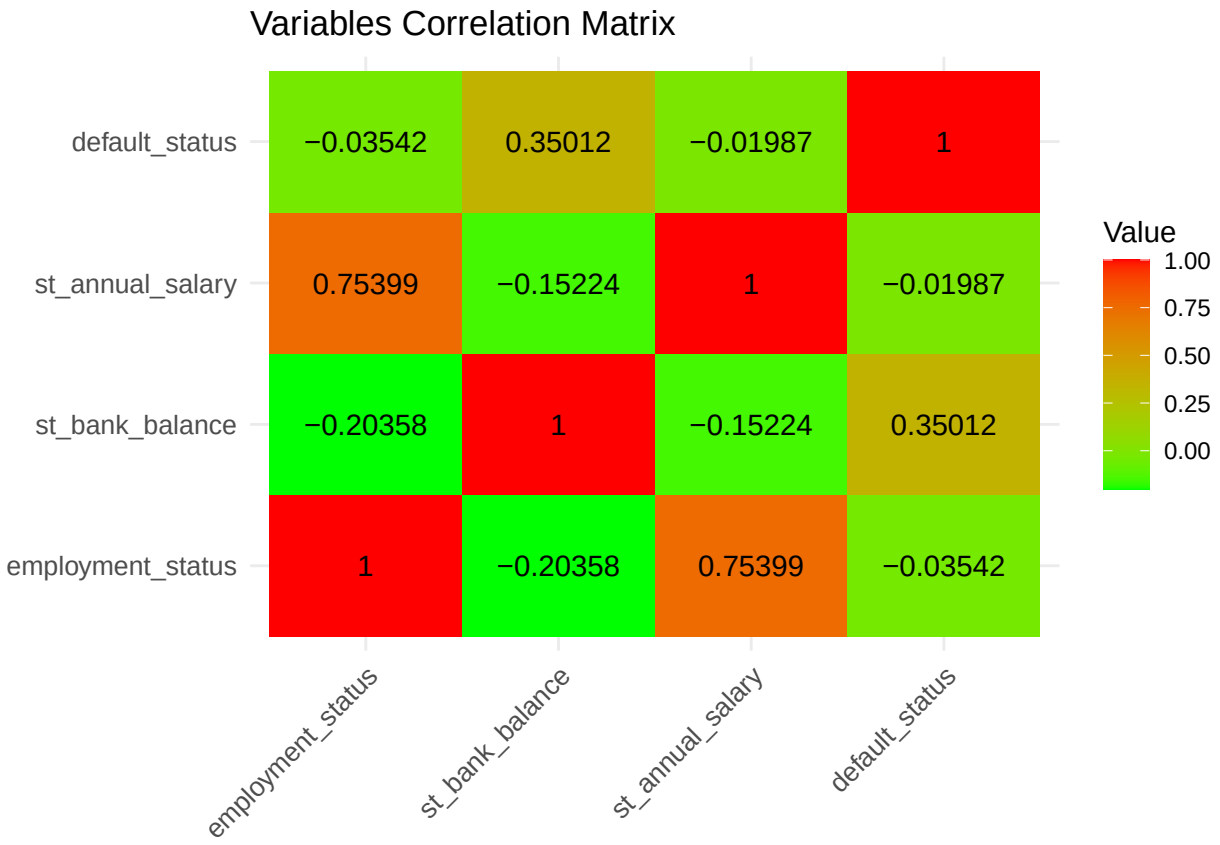




As would be expected, the distributions of the standardised variables are positively skewed since before they were standardised, their values were non-negative.

Correlation between Variables

Analysis of the correlations between predictors and between the response and each predictor will help determine which predictors are worth including in the final fitted logistic regression model.



Of the three predictors, it is clear that the standardised bank balance (st_bank_balance) shares the strongest correlation with default status (default_status), and this correlation is positive as one would expect intuitively. Employment status and annual salary share weak correlation with default status (approximately uncorrelated). Furthermore, employment status and standardised annual salary share a strong positive correlation and it may be worth removing one from the fitted logistic regression model. If either were to be removed, it may make more sense that annual salary were removed as employment status is much more easily measurable.

Modelling

Only 333 of the 10000 observations defaulted; therefore, default is relatively rare. A stratified approach was used to partition the dataset into a training and a validation set to account for the class imbalance in the dataset.

```
logit_model.ebs <- glm(default_status ~ employment_status + st_bank_balance + st_annual_salary,
                       family = binomial,
                       data = T.set)

logit_model.eb <- glm(default_status ~ employment_status + st_bank_balance,
                      family = binomial,
```

```
                      data=T.set)

logit_model.bs <- glm(default_status ~ st_bank_balance + st_annual_salary,
                      family = binomial,
                      data=T.set)

logit_model.b <-glm(default_status ~ st_bank_balance,
                    family = binomial,
                    data=T.set)
```

It is clear that st_bank_balance is the strongest predictor of default_status. There are four combinations of predictors that include st_bank_balance and a logistic regression model using each such combination of predictors was fitted and they will be compared.

E := employment_status, B := st_bank_balance, S := st_annual_salary

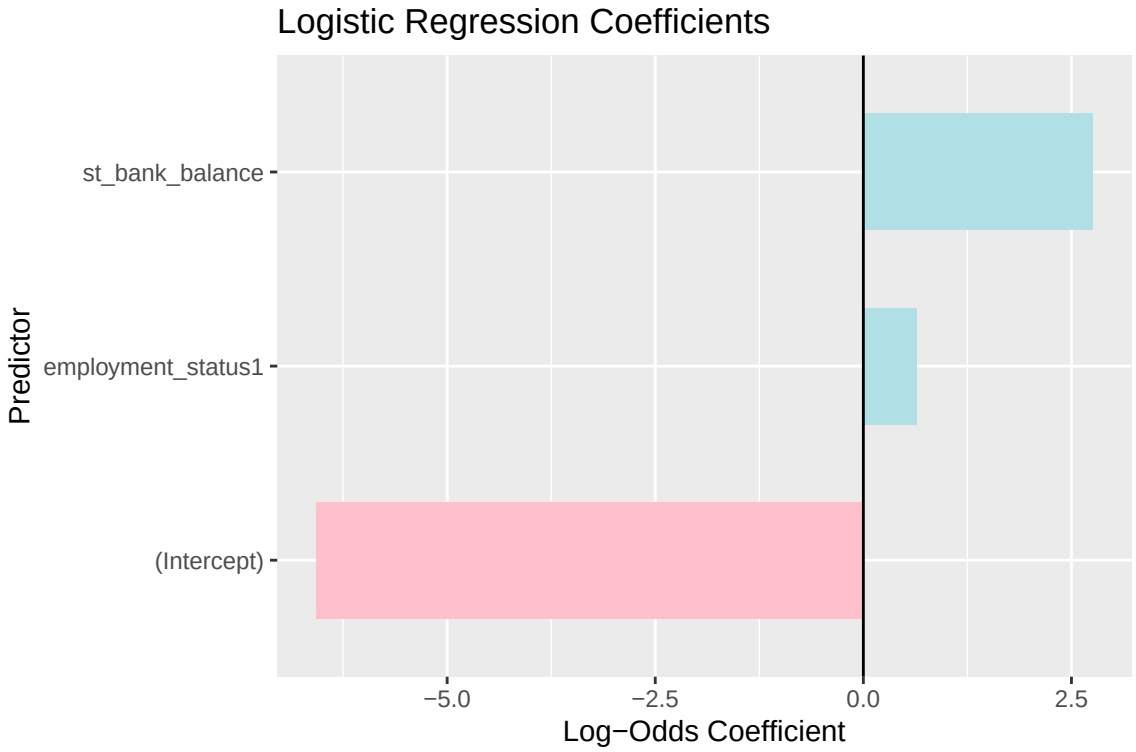
Model Comparison

```
##         Residual_Deviance Null_Deviance      AIC Predictors
## Model 1          1265.615      2340.628 1273.615    E, B, S
## Model 2          1265.616      2340.628 1271.616       E, B
## Model 3          1271.703      2340.628 1277.703       B, S
## Model 4          1282.067      2340.628 1286.067          B
```

Null deviance is the deviance of the intercept-only model and serves as a baseline. It is the same across all four models and therefore lacks discriminative power in this case. Residual deviance measures the unexplained variation once the predictors are included; lower values indicate better fit. AIC is the most useful metric here, as it balances residual deviance with model complexity by applying a penalty for additional predictors. Model 4, which considers only st_bank_balance, is the simplest but has the poorest fit since its AIC value the greatest by a wide margin. Including at least one additional predictor to Model 4 adds predictive power. Model 2 outperforms Model 3 in both AIC and residual deviance and hence Model 2 dominates Model 3. Incorporating st_annual_salary into Model 2 to obtain Model 1 yields a negligible improvement to residual deviance, but worsens AIC. Therefore, the most parsimonious model and best-fitting model is Model 2 (employment_status, st_bank_balance).

Model Evaluation

Model Coefficients



Baseline Probability

In the logistic regression model, the intercept represents the log-odds of the outcome. Here the intercept is large and negative, indicating the baseline probability of default is small. Baseline in this context refers to (employment_status = 0, st_bank_balance = 0). A baseline observation is one that is unemployed and has exactly the mean bank balance of all the observations (since bank balance was standardised).

We can obtain the baseline probability of default by inputting the intercept into the log-odds (intercept) into the logistic transformation f(x) = 1/(1+e^(-x)).

```
baseline_prob.1 <- plogis(logit_model$coefficients[1])
```

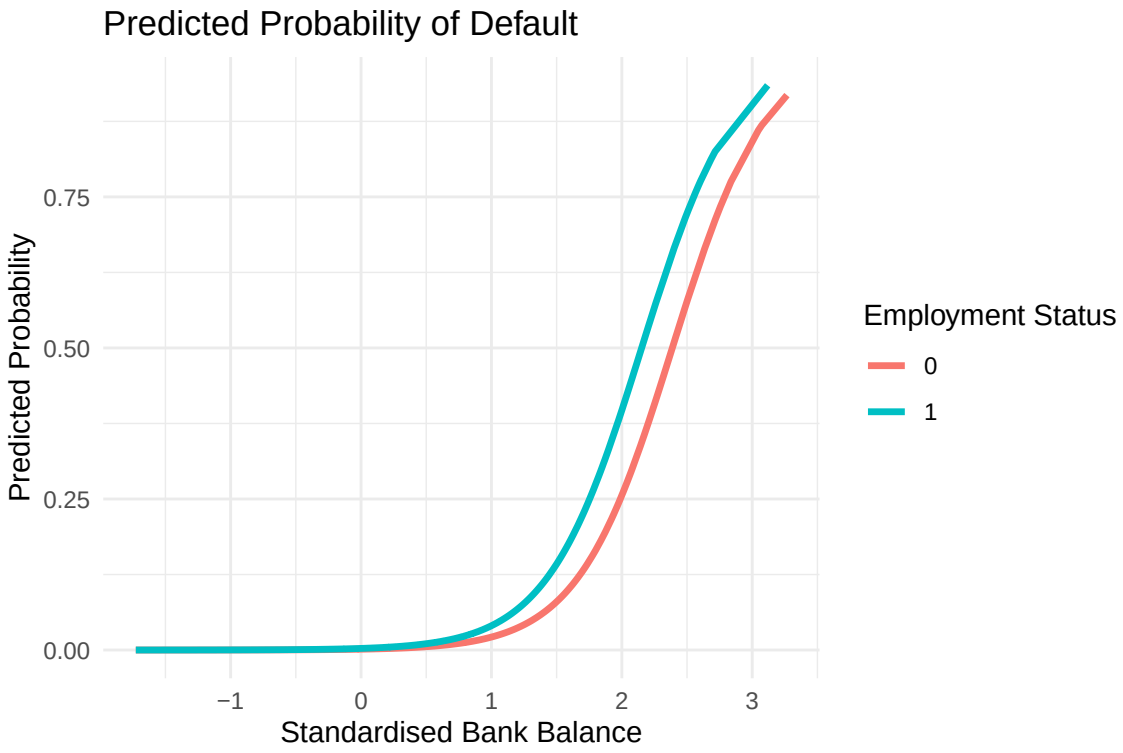
```
## [1] "Baseline probability using logistic transformation on model intercept: 0.00139388843815853"
```

Reproducing the baseline probability by applying the model to an artificial baseline observation (employment_status = 0, st_bank_balance = 0):

```
baseline_obs <- data.frame(employment_status = as.factor(0),
                           st_bank_balance = 0)
baseline_prob.2 <- predict(logit_model, newdata = baseline_obs, type = "response")
```

```
## [1] "Baseline probability obtained by applying model to baseline observation: 0.00139388843815853"
```

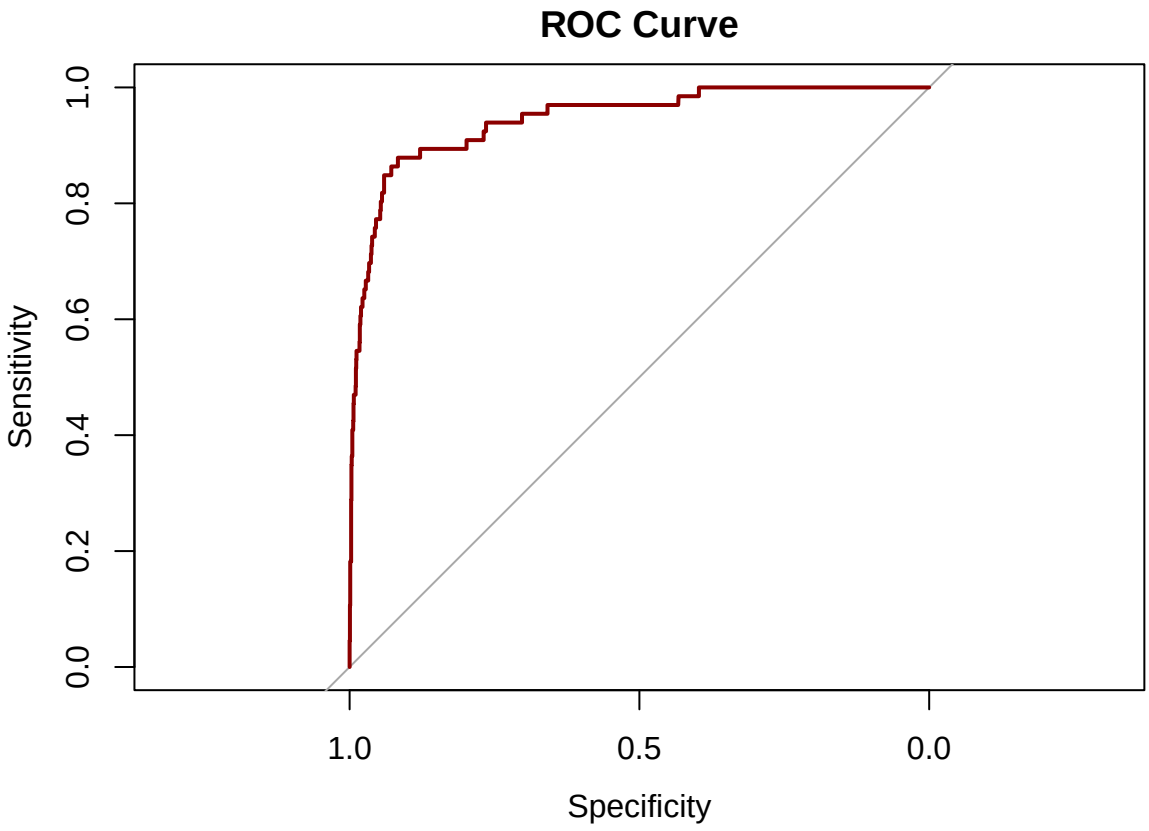
Visualisation of Model



The sigmoid graph above shows the predicted probability of default as a function of the standardised bank balance, with separate curves for each employment status. The upward shape indicates that as standardised bank balance increases, the probability of default also increases, non-linearly. Thus, observations with below-average bank balances have low predicted probability of default, whereas observations with above-average bank balances have high predicted probability of default. At any bank balance value, if the observation is employed, the predicted probability of default is at least as great as if the observation were unemployed. Thus, there is a strong positive non-linear relationship between bank balance and default risk and employment status shifts that level of risk.

Evaluation of the Model

Receiver Operating Characteristic Curve (ROC)




```
## [1] " ROC AUC: 0.945347160168681 Gini: 0.890694320337362"
```

Based on the Area under the Curve (AUC), there is a 94,5% probability that the model will assign a higher predicted probability of default to a randomly selected defaulter than to a randomly selected non-defaulter. A Gini coefficient of 100% indicates perfect predictive accuracy, and this model has a Gini coefficient of 89,07%. Overall, this logistic regression demonstrates a high level of discrimiatory performance.

Confusion Matrix

Ordinarily, when there is a roughly even split in the response variable of a binary classification scenario, a cutoff of 50% is used to produce a confusion matrix. However, in this dataset, the positive class (default_status=1) is relatively rare, with 333 out of 10000 observations. Therefore, a cutoff equal to the percentage of defaulters in the validation set is used instead.

```
print(confusion_matrix)
```

```
##        Predicted
## Actual    0    1
##      0 1660  273
##      1    7   59
```

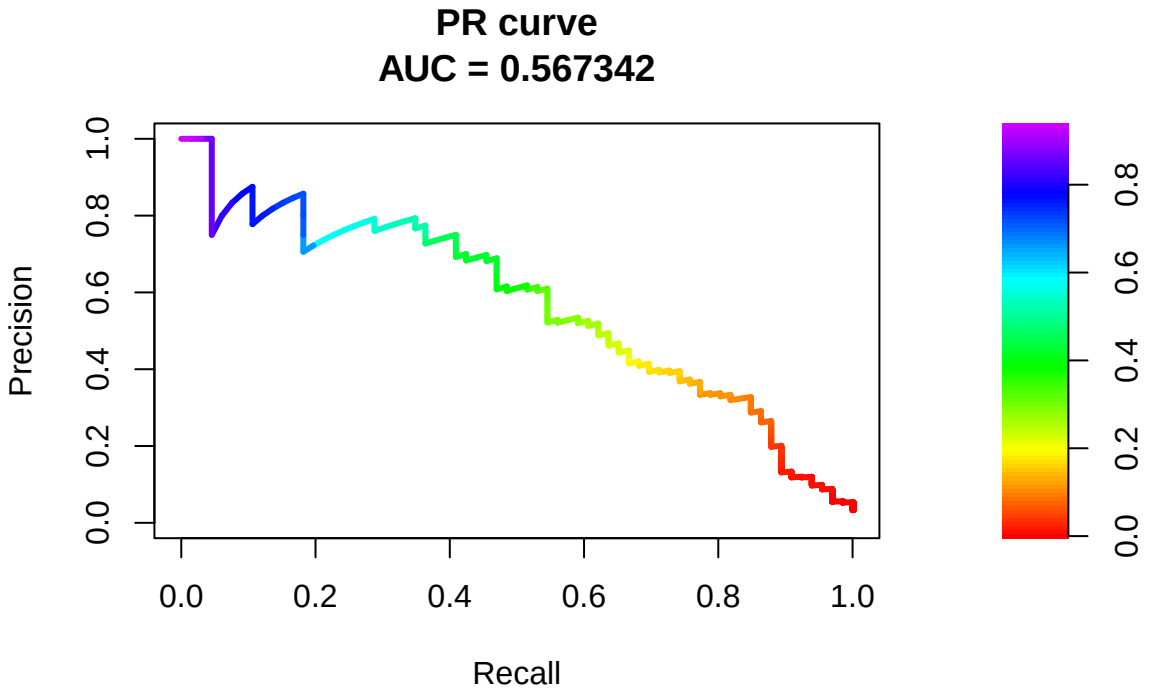
Accuracy, Sensitivity, Specificity and Precision

```
print(performance_metrics)
```

```
##                     Performance Metric
## Accuracy                     0.8599300
## Sensitivity                  0.8939394
## Specificity/Recall           0.8587688
## Precision                    0.1777108
```

Accuracy is how well the model correctly classifies observations (true positives and true negatives). Sensitivity is the ability of the model to detect defaulters and specificity is the ability of the model to identify non-defaulters. Precision is the probability that a predicted defaulter actually defaulted. Each of the accuracy, sensitivity and specificity metrics for this model were strong Precision, however, was low; this is expected due to the class imbalance resulting from the rarity of defaulters in the dataset.

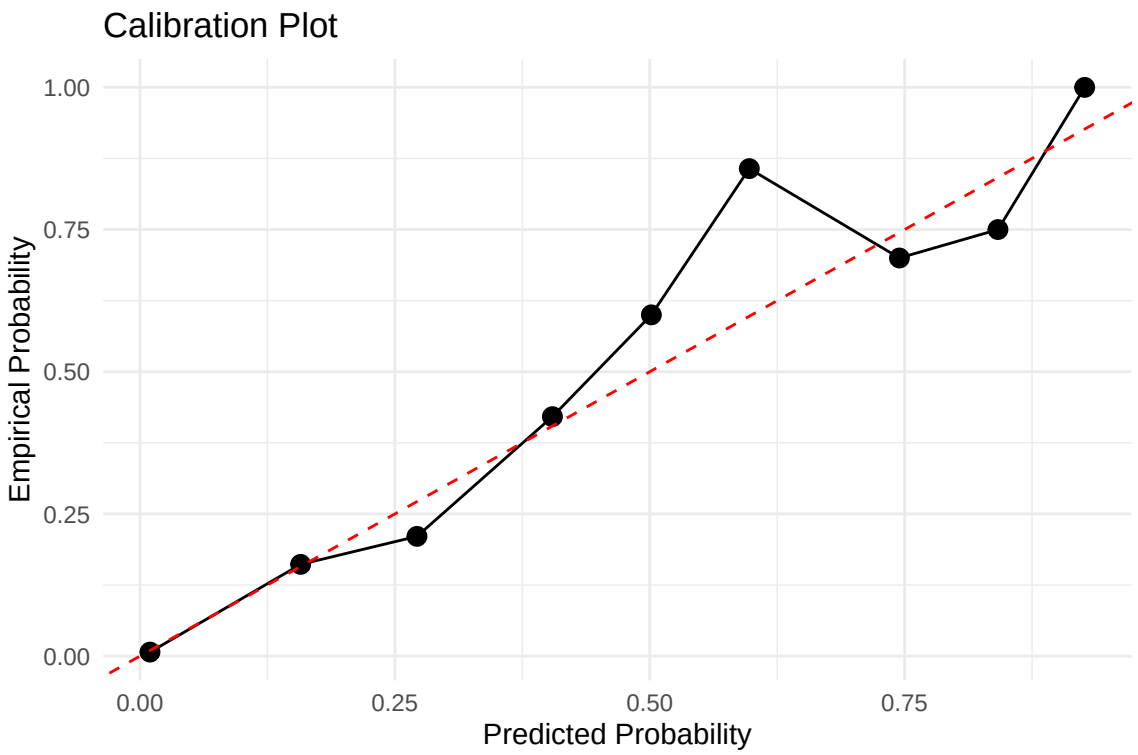
Precision-Recall Curve (PR)



As seen above, the precision of this model is low (0,178), which is to be expected when there is a class imbalance in the response variable - due to the rarity of defaults (3,33%). This has led to a relatively low PR AUC (0,567), however the emphasis on this evaluation measure in this case is low and it is not indicative of a weak model. Instead, this metric complements the high ROC AUC and gives a realistic assessment of the model's practical performance on the minority class (defaulters).

Calibration Plot

A calibration plot is essential for determining how reliably a model's predicted probabilities correspond to actual observed frequencies of the positive class (default_status = 1). The predicted probabilities are obtained by inputting the observations from the validation set into the logistic transformation. The observed values of default status from the validation set are partitioned into n=10 bins and a mean for each bin is computed and used as its observed empirical probability.



This calibration plot adheres relatively closely to the 45-degree reference line, suggesting that the model is reasonably well-calibrated. At low probabilities (0-0,4) observed default probabilities closely match predictions, indicating good calibration. In the mid-range (0,4-0,6) the model shows more fluctuation with some under- and over-prediction across bins, suggesting instability likely due to limited observations over than range. In the upper range of probabilities (0,6-1) there is even more instability as the number of observations of predicted probabilities in that range is even fewer; under- and (especially) over-prediction occur to a greater extent in this upper range. Overall, given that the model only includes employment status and standardised bank balance as predictors, the calibration is strong and no post-processing on the predicted probabilities (e.g. isotonic regression or Platt scaling) is required.

Discussion

The final logistic regression classification model provides important insights into the factors associated with default probability. Most notably, bank balance provided the majority of the signal with employment status providing additional predictive power. Employment status and annual salary exhibited high collinearity and so annual salary was excluded from the final model. This is to the benefit of any potential future use of the model as employment status, being a binary factor, is more easily measurable than annual salary. This selection of predictors led to the most parsimonious, best-fitting and most interpretable model.

The model shows strong discriminatory power and probability calibration indicating that it can reliably differentiate between low-risk and high-risk observations while producing probabilities that align with empirical default frequencies.

Several limitations should be acknowledged. Class imbalance in the response variable likely affected model training and evaluation, potentially resulting in a bias toward to majority class (non-defaulters). The measured performance metrics remain informative, however the minority class (defaulters) may still be underestimated. The model is further limited by the available predictors, e.g. neither credit history for the observations nor a length of study duration were provided. Furthermore, the model's performance is based entirely on the historical patterns on the data and will degrade as underlying risk dynamics inevitably shift.

The model could be improved by incorporating additional predictors that capture latent risk drivers. Further performance improvements could also be achieved through post-processing of the model's predicted probabilities such as isotonic regression or Platt scaling.

Conclusion

This project developed and evaluated a logistic regression model for a binary classification problem, specifically the measure of default risk. It used a structured modelling pipeline including data preprocessing, collinearity assessment, feature selection, model fitting and performance evaluation. The final model achieved a strong balance between accuracy, stability and interpretability. The results indicated that a parsimonious set of features - employment status and bank balance - was effective at capturing default risk, while avoiding highly correlated predictors.

In the risk management context, this model's calibrated probability output can support more informed, data-driven decision-making, such as identifying and prioritising high-risk individuals and improving resource allocation. The model provides predicted probabilities and a credit rating system can be developed to score individuals commensurate to the default risk they pose.

The methods used in this project demonstrate a rigorous approach to risk modelling. The current model provides a defensible baseline, and its effectiveness can be enhanced through a more robust dataset, improved handling of class imbalance, post-processing of predicted probabilities, and continual updating to remain relevant.